Machine Learning Approaches for Predicting Goal Difference and Home Team Loss in Football

Sidak Singh Arora

Abstract

Predicting the outcomes of football matches is a longstanding challenge in sports analytics. In this project, we address two related tasks: (1) predicting the **goal difference** between teams as a regression problem, and (2) predicting whether the **home team will lose** as a classification problem. We use features such as average team attributes, playing styles, aggression, and home advantage to build predictive models. For goal difference prediction, we evaluate the performance of **Linear Regression** and **Random Forest Regressor**. For home team loss prediction, we compare **Logistic Regression** (with **PCA**), **Support Vector Classifiers (SVC)** with linear and RBF kernels, and **Random Forest Classifier**. **PCA** is also applied to assess the impact of dimensionality reduction. Our results demonstrate that ensemble methods such as **Random Forests** are particularly effective for classification, while simple models like **Linear Regression** perform competitively on regression tasks.

1 Introduction

The European Soccer Database on Kaggle contains match data spanning eight seasons (2008–2016), including team lineups, player ratings, and match locations. We begin by preprocessing the data — extracting relevant features from multiple tables, handling missing values, and encoding categorical attributes into numerical formats — to uncover patterns that may enhance predictive modeling.

Following data preparation, we employ Linear Regression and Random Forest Regression to predict match outcomes in terms of goal difference. Model performance is evaluated using Root Mean Square Error (RMSE) and visualized by plotting predicted vs. actual goal differences. Additionally, we compute the R^2 score to assess the proportion of variance explained by each model.

We then approach a classification task: predicting whether the home team will avoid a loss. To do this, we train multiple models — including Linear SVM, RBF SVM, Random Forest Classifier, and Logistic Regression with PCA, varying the number of principal components from 2 to 75 — and compare their classification accuracies. We further apply K-Fold Cross-Validation on Linear SVM, RBF SVM, and Random Forest Classifier to evaluate their consistency across different data splits. For each model, we also tune hyperparameters across a defined range and graph the resulting accuracies to identify the most effective values.

2 Data Preprocessing

2.1 Feature Extraction

The first step in data preprocessing involved selecting the most relevant features for modeling. Key features included the average attributes of the starting 11 players, the number of losses in each team's last five matches, goals scored by the team, and goals conceded over the same period. Additionally, the attacking and defensive work rates, originally represented as categorical strings ("Low", "Medium", "High"), were encoded numerically as 1.0, 2.0, and 3.0 respectively to facilitate model training.

2.2 Data Cleanup

Several matches lacked lineup data, which is critical since the average ratings of the starting 11 are among the most influential features in predicting both the winner and goal difference. Consequently, matches without complete lineup information were excluded from the dataset.
In addition, some players had missing values for their attacking and defensive work rates. To address this, missing entries were imputed with a work rate of "Average," which represents the median category across all players for both attributes.

2.3 Train Test Validation Split

To evaluate the model's performance, the dataset was split into 80% for training and 20% for testing. During model training for home team loss prediction, 5-fold cross-validation was employed to ensure robustness and reduce the risk of overfitting, allowing the model to be validated across multiple subsets of the training data.

3 Methodology

3.1 Linear Regression to predict goal difference

Linear Regression performs well when a linear relationship exists between the input features and the target variable. However, it struggles to capture non-linear patterns in the data. In this case, the model achieves a Mean Squared Error (MSE) of 2.81. Since MSE squares the differences between predicted and actual values, the Root Mean Squared Error (RMSE) provides a more interpretable metric in the original units. The RMSE is 1.67, indicating that the predicted goal difference is, on average, off by approximately 1.67 goals. The coefficient of determination, R^2, is 0.20, suggesting that the model explains only 20% of the variance in the target variable and fails to adequately capture the underlying structure of the data.

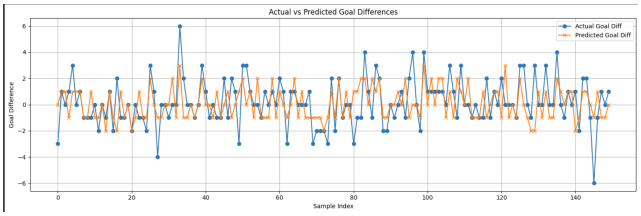

Figure 1: Linear Regression and Test Set Plot

3.2 Random Forest Regressor to predict goal difference

Random Forests are robust to non-linear relationships and tend to perform well on high-dimensional datasets. In this case, the model achieves a Mean Squared Error (MSE) of 2.79. Since MSE squares the prediction errors, the Root Mean Squared Error (RMSE) is often preferred for interpretability, as it represents the average prediction error in the same units as the target variable. The RMSE here is 1.67, meaning the model's goal difference predictions are, on average, off by approximately 1.67 goals. The coefficient of determination, $R^2 = 0.19$, indicates that the model explains only 19% of the variance in the target variable, suggesting that either the features used are not strongly predictive or the outcome is inherently difficult to predict (Figure 2).

3.3 Linear SVC to predict home team loss

Linear Support Vector Classifier (SVC) attempts to find a decision boundary that maximizes the margin between the closest data points from each class—these points are known as support vectors. It performs well on linearly separable datasets, especially in lower-dimensional spaces. However, its performance tends to degrade on datasets with highly non-linear relationships unless additional preprocessing or kernel transformations are applied.

In our case, without cross-validation, the Linear SVC achieved an accuracy of 73.2% at $C = 10^{-4}$.

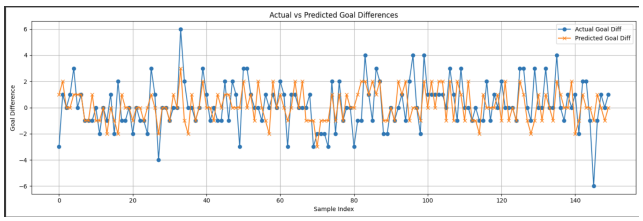

Figure 2: Random Forest Regression and Test Set Plot

When evaluated using cross-validation, the model achieved an accuracy of 73.1%. Refer to Figure 3 for the plot of Accuracy vs. C.

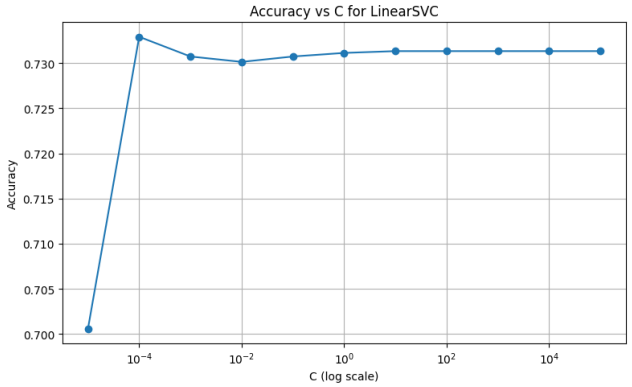

Figure 3: Accuracy Vs C Plot

3.4 Random Forest Classifier to predict home team loss

Random Forests are highly robust to outliers and effective at identifying relevant features, even in high-dimensional datasets. Without cross-validation, the model achieved an accuracy of 73.4% using 500 decision trees. When evaluated using 5-fold cross-validation, the model achieved an accuracy of 72.9%.

Refer to Figure 4 for the plot of Accuracy vs. Number of Trees.

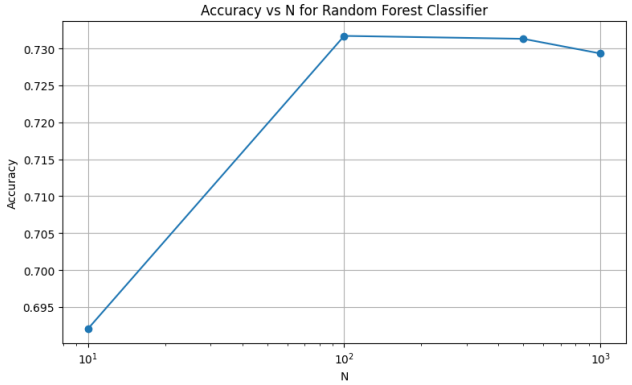

Figure 4: Accuracy Vs Number of trees Plot

3.5 Logistic Regression with PCA to predict home team loss

PCA reduces the dimensionality of the data by projecting it onto a linear combination of features that capture the most variance. It works best with linearly distributed data. We applied PCA to reduce the input features before training and testing with Logistic Regression, which performs well on linearly separable data. Logistic Regression is also computationally efficient and provides interpretable results. Using Logistic Regression with varying numbers of PCA components, we achieved a maximum accuracy of 73.3%. Refer to Figure 5 for the plot of Logistic Regression Accuracy vs. PCA Components.

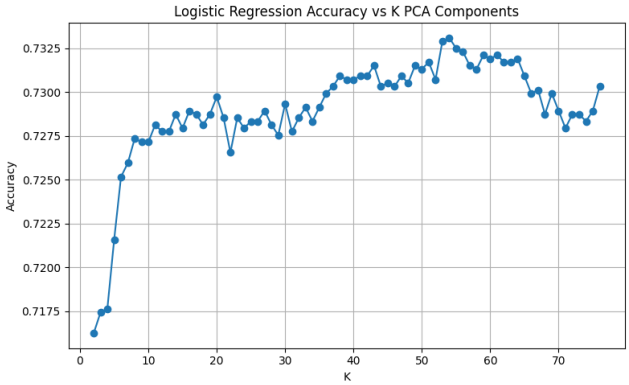

Figure 5: Logistic Regression Accuracy vs. PCA Components

3.6 RBF SVM to predict home team loss

The RBF SVM maps input data into a higher-dimensional space using a kernel function, allowing it to find a non-linear decision boundary that separates the classes more effectively.

We achieved an accuracy of 70.0% using this model. By applying 5-fold cross-validation, we improved the accuracy to 71.2%.

Refer to Figure 6 for the plot of RBF SVM Accuracy vs. Gamma values.

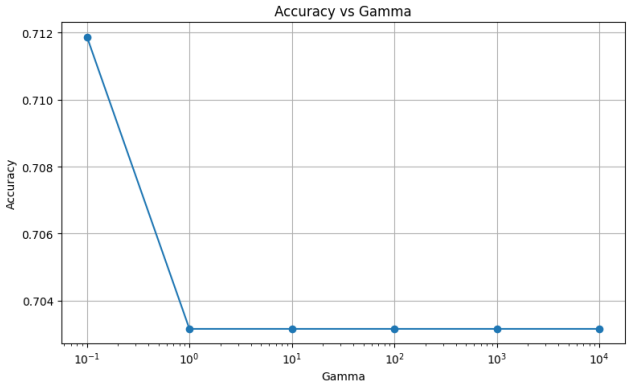

Figure 6: RBF SVM Accuracy vs. Gamma values

4 Discussion

To predict goal difference, both the linear regressor and the random forest regressor performed similarly, suggesting a predominantly linear relationship between the input features and the target. Both models had an RMSE of approximately 1.67 goals per game, indicating they capture some underlying patterns but still have notable prediction errors.

Since RMSE penalizes large errors more heavily, outliers—such as high-scoring games—significantly contribute to the overall error. We also observed that both models tend to make conservative predictions clustered around the mean goal difference of 2 goals. This behavior aligns with the distribution of the dataset, where most games end with a goal difference close to 2.

However, the models struggle with games that have a goal difference greater than 4, leading to larger errors. These cases are relatively rare, so the models' predictions are biased toward the more frequent, average scenarios.

We also observe from the R^2 score that both models were only able to capture 19% of the underlying variance in goal difference. This indicates a weak relationship between the input features and the target variable. This result is reasonable, as goal difference can vary significantly due to unpredictable factors such as defender errors, individual brilliance, or game-day conditions—none of which are captured in the current feature set.

To predict whether the home team loses, all models performed similarly, with the best accuracy of 73.3% achieved by PCA-reduced Logistic Regression. However, all models hovered around the 73% mark. Logistic Regression being the top performer supports the assumption of linearity between input features and the target, as also suggested by the goal difference prediction task. This implies that a linear combination of features like the playing 11's average stats, recent performance, and goals scored is indicative of whether a team will lose.

That said, an accuracy of 73% also highlights the inherent unpredictability in football outcomes. This is understandable, as factors like player errors, match-day conditions, and mental or physical states are difficult to quantify and are not captured in the data used.

A notable example of such unpredictability was when FC Barcelona lost to Celta Vigo (4-1), despite being the clear favorite, coming off five unbeaten matches, and fielding their strongest lineup.

5 Limitations

One limitation of the model is the absence of certain influential features in the dataset, such as game-day conditions or the head-to-head history between team coaches—both of which could impact the outcome. The dataset did not include these attributes, which may contribute to reduced accuracy.

Another challenge is the high dimensionality of the dataset. A more granular comparison—such as evaluating individual matchups between each team's attackers, midfielders, and defenders—could improve prediction accuracy. However, this would significantly increase model complexity and training time, making it less feasible for our scope.

6 Conclusion

In this report, we explored various models for predicting goal difference and determining whether the home team would lose. Our analysis revealed some degree of linear dependence between input features and target outcomes. While the models demonstrated moderate performance, they were able to capture underlying patterns in the data.

We also discussed key limitations of the dataset, such as missing contextual features and high dimensionality, which likely impacted model accuracy. Despite these constraints, the results are promising and suggest that further refinement and feature engineering could lead to improved predictive performance.

References

[1] Mathien, H. (2016). European Soccer Database. In Kaggle Datasets. Available at: https://www.kaggle.com/datasets/hugomathien/soccer